



Heriot-Watt University
Research Gateway

Systematically corrupting data to assess data linkage quality

Citation for published version:

Alsadeeqi, A, Gray, AJG, Christen, P, Akgün, Ö & Dalton, T 2017, 'Systematically corrupting data to assess data linkage quality', The UK Administrative Data Research Network Annual Research Conference 2017, Edinburgh, United Kingdom, 1/06/17 - 2/06/17.

Link:

[Link to publication record in Heriot-Watt Research Portal](#)

General rights

Copyright for the publications made accessible via Heriot-Watt Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

Heriot-Watt University has made every reasonable effort to ensure that the content in Heriot-Watt Research Portal complies with UK legislation. If you believe that the public display of this file breaches copyright please contact open.access@hw.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Systematically corrupting data to assess data linkage quality

Ahmad Alsadeeq¹, Alasdair J G Gray¹, Peter Christen², Özgür Akgün³, Tom Dalton³

¹School of Mathematical and Computer Sciences, Heriot-Watt University, UK;

²Research School of Computer Science, The Australian National University, Canberra, Australia; ³School of Computer Science, University of St. Andrews

Various algorithms have been developed to automatically link historical records based on a variety of string matching techniques. These generate an assessment of how likely two records are to be the similar. However, it remains unclear how to assess the quality of the linkages computed due to the absence of absolute knowledge of the correct linkage of real historical records – the ground truth. The creation of synthetically generated datasets for which the ground truth linkage is known to help with the assessment of linkage algorithms but the data generated is commonly too clean to be representative of historical records.

We are interested in assessing record linkage algorithms under different data quality scenarios, e.g. with errors typically introduced by a transcription process or where books can be nibbled by mice. We are developing a data corrupting model that injects corruptions into datasets based on given corruption methods and probabilities. We have classified different forms of corruptions found in historical records into four types based on the effect scope of the corruption. Those types are character level (e.g. an 'f' is represented as an 's' - OCR Corruptions), attribute level (e.g. gender swap - male changed to female due to false entry), record level (e.g. missing records due to different reasons like loss of certificate), and group of records level (e.g. lost parish records in fire). This will give us the ability to evaluate record linkage algorithms over synthetically generated datasets with known ground truth and with data corruptions matching a given profile. In this paper, we describe in detail these four types of corruptions and corresponding examples.